Scene stitching with event-driven sensors on a robot head platform

Klein, P ; Conradt, J ; Liu, S-C

Abstract: This paper describes a robot head platform which holds a pair of event-based Dynamic Vision Sensor (DVS) retinas and microphones connected to an event-based binaural AEREAR2 VLSI cochlea system. The platform has 6 degrees of freedom (DOF): 2 for the neck, and 2 for each of the DVS retinas. Two applications using this platform are described: the first is image stitching of a scene larger than the field of view of the individual retinas as the head pans and tilts and the second is selective image painting of the local visual scene around spatially displaced sound sources. This platform allows for the investigation of event-driven sensory-action models that use the information from multiple event-based sensor modalities in real-time scenarios.

DOI: https://doi.org/10.1109/ISCAS.2015.7169173

Scene Stitching with Event-Driven Sensors on a Robot Head Platform

Philipp Klein[1], Jorg Conradt[2], and Shih-Chii Liu[1]
[1]Institute of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland
[2]Technische Universität München, München, Germany

Abstract—**This paper describes a robot head platform which holds a pair of event-based Dynamic Vision Sensor (DVS) retinas and microphones connected to an event-based binaural AEREAR2 VLSI cochlea system. The platform has 6 degrees of freedom (DOF): 2 for the neck, and 2 for each of the DVS retinas. Two applications using this platform are described: the first is image stitching of a scene larger than the field of view of the individual retinas as the head pans and tilts and the second is selective image painting of the local visual scene around spatially displaced sound sources. This platform allows for the investigation of event-driven sensory-action models that use the information from multiple event-based sensor modalities in real-time scenarios.**

I. Introduction

Event-driven sensors such as Dynamic Vision Sensors (**DVS**s) [1], and the event-driven binaural AEREAR2 cochlea [2]produce asynchronous spike events in response to visual and auditory stimuli in the world. These sensors have been used in sensory-motor setups to illustrate the efficiency of data-driven computation from such sensors as compared to frame-based sensors [3]. However most of these applications are demonstrated with stationary sensors. The use of the DVS retinas on a moving robotic head with the degrees of freedom to approximate human-like head and eye movements is rare [4], [5]. The inclusion of event-based silicon cochleas with a silicon retina on a robotic platform is also rare [6].

This paper describes such a platform with both types of sensors. The platform consists of a custom robot head with 6 degrees of freedom to approximate human-like head and eye movements. It holds a pair of event-based DVS retinas and microphones going to the binaural AEREAR2 board. The setup is described in Section II followed by descriptions of two examples of how this platform is used for visual and visual-auditory tasks in prototypical real-world scenarios in Sections III and IV.

II. Setup

The robotic head (Fig. 1) consists of the base plate, the camera plate, and a total of six servos (HiTec Digital Robot Servo HSR-5498SG, with housing exchanged for HSR-6980SG for easier mounting). These servos were chosen because they offer a high movement speed with low noise emission in stable positions. To fit the servos to the head, they were placed in another case. The servos are controlled using two NXP LPC1769 microcontrollers over an FDTI (FTDI232HL single channel, 4Mbps data rate) virtual serial connection with a USB interface to the PC. That way, it is possible to set each servo to a certain position using Pulse Width Modulation (PWM) values. Two of the six servos control the neck; the remaining four servos control the DVS retinas.

The base and camera plate are 3D-printed on a Stratasys Dimension 1200es. This production method results in high stiffness, to ensure high accuracy of the mechanics in order to avoid as much vibration as possible. The base plate is attached onto a servo that is inside the neck which allows pan movements around the center axis of the head. The camera plate is connected to the base plate on the sides with bearings and with a servo in the middle of the base plate which allows tilt movements of the camera plate. 3D printing requires about 24h.

The two DVS retinas act as the eyes of the head. Each retina is controlled by 2 independent servos allowing it to pan and tilt independently of the other retina and also of the neck. The retinas are mounted 19.2cm apart from each other and 7.5cm above the base plate.

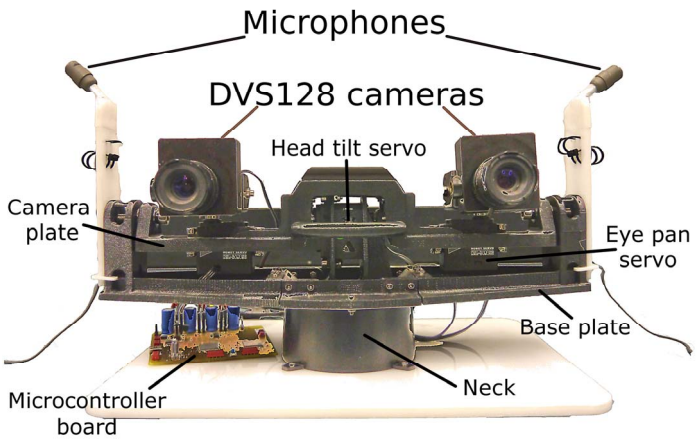


Fig. 1 Robot head with a pair of DVS retinas and microphones. The microphones are connected to the AEREAR2 board shown in Fig. 2.

The two ears receive input from a microphone attached on each side of the robot head, which are spaced by 36.8cm. The microphones are connected to binaural AEREAR2 cochlea PCB (Fig. 2). Because the microphones are connected only to the base plate, they move according to the pan movements of the head.

The DVS retina has a resolution of 128x128 pixels. It produces events (spikes addresses) only if a pixel senses local brightness changes. The pixels output ON and OFF events which code both positive and negative changes in log intensities respectively. Both retinas are connected to the PC through separate USB connections.

The AEREAR2 PCB [2] holds a custom AEREAR2 binaural cochlea chip, digital chips to handle the communication with the cochlea and to transmit the events and the timestamps over USB to the PC. Each cochlea is modeled by a 64-stage cascaded filter bank followed by a half-wave rectifier which models the inner hair cell and an integrate-fire neuron model which models the spiral ganglion cells.

The interfacing of the DVS retinas and the AEREAR2 cochlea sensor, as well as the control of the servos is done using jAER, the open-source Java-based project that allows event-driven processing with address-event representation (AER) systems on PCs [7]. Through this software, it is possible to combine the sensory information from the different sensors and to control the pan and tilt movements of the different motors. The source code for the interaction between the DVS cameras, the cochlea sensor and the motor control is in the jAER package *ch.unizh.ini.jaer.projects.robothead6DOF*. The main class that combines the two different sensors and the robot head control is *ITDImageCreator*. The robot head control is written as a modular set of jAER filters that operate on the sensor output.

1) Robot head control

The robot head is interfaced to the computer using a USB connection and a virtual serial connection on the microcontroller board. Because each servo has a defined address in the microcontroller, it is possible to control each servo individually. The control commands for the microcontrollers consist of a string that contains the address of a specific servo and a PWM value. Based on these control commands the microcontroller drives a specific servo to the new position corresponding to the PWM value. Overall control of the gaze direction and stereo vergence are encapsulated in the class *Head6DOF_ServoController*

2) Scene stitching

To display the histogram image of the scene around the robot head, the DVS spikes are first filtered by the jAER *BackgroundActivityFilter* and then added to the appropriate location within the image frame through the *ImageCreator* filter. The size of this frame depends on the pan angle of the head and the resulting image shifts on the DVS retinas. In this setup the frame has a size of 628x418 pixels when the DVSs have 8mm lenses corresponding to a horizontal field of view (FOV) of 150° and a vertical FOV of 90°. Before the stitching is initialized, the frame is zeroed as represented by a gray value. Every ON event increases the intensity value of a specific pixel towards pure white and each OFF event decreases the intensity value towards pure black. To stitch the histogram image over multiple views, the mapping between the movements of the neck and corresponding shift of the DVS image at each of these positions was calibrated as described in Section III.

3) Determining the position of a sound source

To determine the location of a sound, we use the interaural time difference (**ITD**) that describes the difference in the arrival times of the sound at the 2 ears. The *ITDFilter_robothead6DOF* filter used here is the same filter as described in [8] and was slightly modified for use on this robotic setup. It computes the ITD values in a range of +-800 us to match approximately the chosen range of pan movements of the experiment in Section IV.

4) Painting a histogram image around a sound source

The *ITDImageCreator* filter combines all the filters in (1) to (3) and can control small circular movements of the eyes to generate DVS output when the eyes are otherwise stationary. Additionally it receives information of the current ITD value and can send commands according to this information to the robot head. Lastly it starts the histogram image acquisition. This way it is possible to control the scene stitching as well as the acquisition of the local audio-visual image within a single jAER filter.

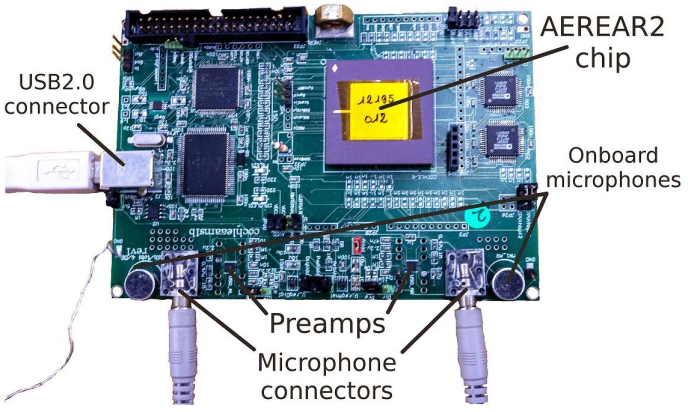


Fig. 2 AEREAR2 PCB and microphone connectors.

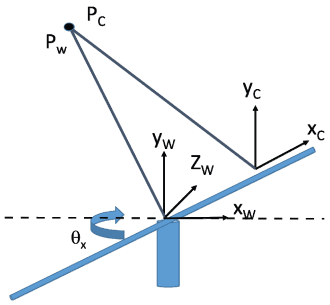
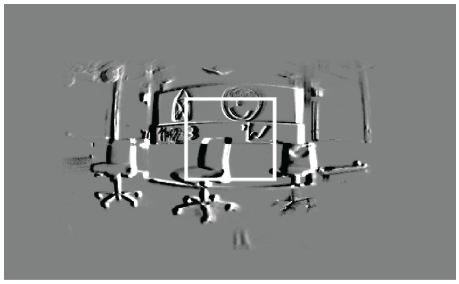
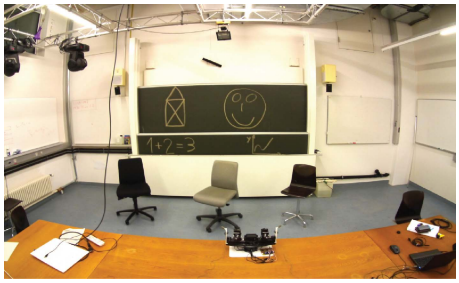


Fig. 3 Mapping of point in world coordinates (P_W) to DVS camera coordinates (P_C).



(a)



(b)

Fig. 4 Comparison between (a) the histogram image of the right DVS retina with an 8mm lens and (b) the image of a conventional camera with a 9mm fisheye lens. In (a), gray means no spikes, white means ON contrast spikes. Black means OFF spikes. Brightest/darkest values=100 spikes. The FOV of the DVS retina (H: 35 deg, V: 26 deg) is indicated by the white rectangle.

III. Scene Stitching

The two experiments described next show the use of the different motorized parts of the robotic head together with the two sensor modalities. The first experiment is the creation of an extended image (I_{EI}) by stitching together the image generated by the DVS spikes (I_{DVS}) by using the pan and tilt movement of the neck. A point in the DVS camera coordinates (P_C) can be transformed to a point in the world (P_W) as shown in Fig. 3 by using Euler angles. Instead of determining the translational and rotational matrices needed for this transformation, in this paper we followed a simpler approach sufficient for this experiment. The pixel coordinate relation between I_{EI} and I_{DVS} is calibrated using a panel of 8x10 LEDs spaced 4cm apart at a distance of 245cm away from the robot head and measuring the movement of the LEDs caused by pan and tilt movements. At this scene distance, the DVS coordinate shift $p_y(\theta_y)$ caused by a tilt θ_y in degrees is $p_y(\theta_y) = 3.4\theta_y$ pixels. and the DVS coordinate shift $p_x(\theta_x)$ caused by a head pan θ_x in degrees is $p_x(\theta_x) = -0.0065 * \theta_x + 3.78$ pixels. The I_{EI} image is thus assembled from DVS histograms according to Eq. (1), where (x,y) are the DVS coordinates:

$$I_{EI}(x1, y1) = I_{DVS}(x, y)$$
$$\text{where } x1 = \lfloor x + p(\theta_x)(\theta_x + \theta_{x\max}) \rfloor, y1 = \lfloor y + p(\theta_y)(\theta_y + \theta_{y\max}) \rfloor \quad (1)$$

and $\theta_{x\max} = 45.5°; \theta_{y\max} = 25°; \theta_x = [-45.5°, 45.5°]; \theta_y = [-25°, 25°]$.

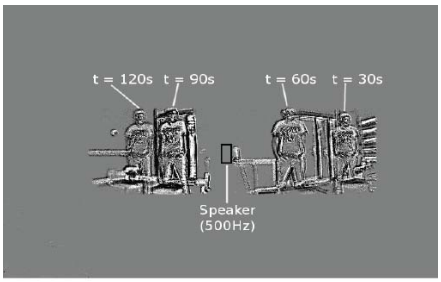
The stitched histogram image in Fig. 4(a) is generated by panning from left to right for various tilt positions. The field of view of the DVS is shown by the white square. The image taken by a conventional camera is shown in Fig. 4(b). A new position command is sent every 50ms and because the steps were only 0.5°, the image required 200s to assemble. The step interval was limited by the microcontroller and the response time of the motors.

IV. Audio-Visual Scene Painting

The second experiment shows how the outputs of two event-based sensor modalities can be combined to imitate the scenario of an auditory source acting as an attentional signal and directing the eyes to move to a particular location in space, while ignoring other sound sources. The location of a sound source on this setup is determined by the ITD algorithm described in [8]. A running histogram of ITDs is computed from the incoming spike events and the histogram peak ITD in microseconds is mapped to θ_x using the following equation after calibration: $\theta_x = (0.0616 * ITD - 0.0288)$ deg. For this experiment the head is restricted to only pan movements. The ITD is computed from channels not stimulated by the motor noise. The maximum pan rate of the head in response to the ITD is about 185deg/s.

Once the head faces the sound source, the DVS retina performs a small circular clockwise movement using the two motors of each eye. This movement is an approximation of the natural microsaccadic movements of the eye of about 3Hz. Each rotation takes about 380ms.

The selective painting of the histogram image around a sound source of interest is also illustrated in this experiment. Two spatially displaced sound sources were used; the first source was a fixed 500 Hz tone from a loudspeaker in the middle of the



(a)

(b)

Fig. 5 Selective visual painting around a sound of interest using a DVS with an 8mm lens. The head only paints the histogram image (a) around the human speaker and ignores the 500Hz pure tone. Brightest/darkest values=10 spikes. The black box shows the position of the loudspeaker. (b) Picture from a conventional camera. FOV of camera is smaller than overall FOV of robot head.

view and the second source was a moving human. The source of interest is the human speaker who walks across the room. The 500Hz tone source was ignored. The sound sources were at a distance of 175cm from the robot head. The human moved to four positions separated by 75cm (Fig. 5 (b)).

The tone and human sound sources were alternately active for 30sec, starting with a 30s speaker tone. Using the ITD algorithm, the head orientated towards the human sound. The human speaker direction was detected by looking for activity in cochlea channels not stimulated by the 500Hz source. While the loudspeaker was active, the human speaker moved to a different position in the room. The resulting image in Fig. 5 (a) after 120s, shows that the scene was only painted when the human voice is present, because only a small part of the loudspeaker on the edge of one of the human fixation views is present in the image.

V. Discussion

This robot head platform is useful for exploring event-driven models using outputs of event-based multi-modal visual and auditory sensors. It will be used for future applications such as object tracking in real-time where the DVSs sit on a moving head rather than a stationary platform. Although the experiments are somewhat simplified, in particular for the last experiment involving selective image painting driven by a desired sound source, our intention is to replace these sound sources with multiple simultaneous human speakers. Testing of more sophisticated algorithms on the DVS output such as the one used on the panoramic DVS setup [9] and the recent super-resolution mosaicing algorithm of [10] can be tested on this platform. With recent work in combining both an IMU and DVS together to stabilize DVS output against camera rotation [11], this platform can be used to investigate event-based models for smooth pursuit, object tracking, and stereo vision.

Acknowledgements

The authors acknowledge T. Delbruck for the support of the filter software in creating a stitched image and H. Finger for the initial ITD filter control of a 2DOF pan-tilt system. This work is partially funded by the EU project SeeBetter (FP7-ICT-270324).

References

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120 dB 15 µs Latency Asynchronous Temporal Contrast Vision Sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous Binaural Spatial Audition Sensor with 2x64x4 Channel Output," *IEEE Trans Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453 – 464, 2013.

[3] J. Conradt, M. Cook, R. Berner, P. Lichtsteiner, R. J. Douglas, and T. Delbruck, "A Pencil Balancing Robot Using a Pair of AER Dynamic Vision Sensors," in *IEEE International Symposium on Circuits and Systems (ISCAS) 2009*, Taipei, 2009, pp. 781–784.

[4] C. Bartolozzi, F. Rea, C. Clercq, M. Hofstatter, D. B. Fasnacht, G. Indiveri, and G. Metta, "Embedded neuromorphic vision for humanoid robots," in *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011, pp. 129–135.

[5] F. Rea, G. Metta, and C. Bartolozzi, "Event-driven visual attention for the humanoid robot iCub," *Neuromorphic Eng.*, vol. 7, p. 234, 2013.

[6] V. Chan, C. T. Jin, and A. van Schaik, "Neuromorphic Audio-Visual Sensor Fusion on a Sound-Localizing Robot," *Front Neurosci*, vol. 6, pp. 1–9, 2012.

[7] "jAER Open Source Project," *jAER Open Source Project*. [Online]. Available: http://jaerproject.org. [Accessed: 17-Sep-2013].

[8] H. Finger and S.-C. Liu, "Estimating the location of a sound source with a spike-timing localization algorithm," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2011, pp. 2461–2464.

[9] A. N. Belbachir, S. Schraml, M. Mayerhofer, and M. Hofstatter, "A Novel HDR Depth Camera for Real-Time 3D 360° Panoramic Vision," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 425–432.

[10] H. Kim, A. Handa, R. Benosman, S.-H. Ieng, and A. J. Davison, "Simultaneous Mosaicing and Tracking with an Event Camera," in *Proceedings of the British Machine Vision Conference (BMVC)*, Nottingham, 2014.

[11] T. Delbruck, V. Villaneuva, and L. Longinotti, "Integration of Dynamic Vision Sensor with Inertial Measurement Unit for Electronically Stabilized Event-Based Vision," in *Proc. 2014 Intl. Symp. Circuits and Systems (ISCAS 2014)*, Melbourne, Australia, 2014, pp. 2636–2639.